Scatter Diagrams and Correlation

Gene Quinn

Bivariate Data

In this section we deal with data containing **two** variables.

Bivariate Data

In this section we deal with data containing **two** variables.

Usually this means we have a sample where we record *two* measures on each subject in the sample.

Bivariate Data

In this section we deal with data containing **two** variables.

Usually this means we have a sample where we record *two* measures on each subject in the sample.

The term **bivariate data** is used to describe data of this kind.

Bivariate Data

In this section we deal with data containing **two** variables.

Usually this means we have a sample where we record *two* measures on each subject in the sample.

The term **bivariate data** is used to describe data of this kind.

For example, we might record the barometric pressure and sustained wind speed of a hurricane at different hours:

Time	Pressure (mb)	Sustained wind speed
1 AM	980	85
4 AM	970	92
7 AM	950	110

Bivariate Data

Often we are interested in using one variable to predict the value of the other.

Bivariate Data

Often we are interested in using one variable to predict the value of the other.

Often, this is because one variable is easier to measure than the other.

Bivariate Data

Often we are interested in using one variable to predict the value of the other.

Often, this is because one variable is easier to measure than the other.

In the case of a hurricane, a barometer can be dropped from an airplane to measure the barometric pressure.

Bivariate Data

Often we are interested in using one variable to predict the value of the other.

Often, this is because one variable is easier to measure than the other.

In the case of a hurricane, a barometer can be dropped from an airplane to measure the barometric pressure.

Once we have the barametric pressure, historical records allow us to predict the wind speed.

Bivariate Data

Often we are interested in using one variable to predict the value of the other.

Often, this is because one variable is easier to measure than the other.

In the case of a hurricane, a barometer can be dropped from an airplane to measure the barometric pressure.

Once we have the barametric pressure, historical records allow us to predict the wind speed.

If barometric pressure can be measured more precisely than wind speed, which varies considerably due to gusts, we can improve the accuracy of our wind speed estimate.

Bivariate Data

In a situation where one variable (barometric pressure) can be used to explain or predict another variable (wind speed), the variable whose value we wish to explain is called the **response** variable.

Bivariate Data

In a situation where one variable (barometric pressure) can be used to explain or predict another variable (wind speed), the variable whose value we wish to explain is called the **response** variable.

In the preceding example, wind speed would probably be considered the response variable.

Bivariate Data

In a situation where one variable (barometric pressure) can be used to explain or predict another variable (wind speed), the variable whose value we wish to explain is called the **response** variable.

In the preceding example, wind speed would probably be considered the response variable.

The variable we will use to explain the response variable is called the **predictor** or **explanatory** variable.

Bivariate Data

In a situation where one variable (barometric pressure) can be used to explain or predict another variable (wind speed), the variable whose value we wish to explain is called the **response** variable.

In the preceding example, wind speed would probably be considered the response variable.

The variable we will use to explain the response variable is called the **predictor** or **explanatory** variable.

In the preceding example, barometric pressure would probably be considered the predictor variable.

Bivariate Data

In a situation where one variable (barometric pressure) can be used to explain or predict another variable (wind speed), the variable whose value we wish to explain is called the **response** variable.

In the preceding example, wind speed would probably be considered the response variable.

The variable we will use to explain the response variable is called the **predictor** or **explanatory** variable.

In the preceding example, barometric pressure would probably be considered the predictor variable.

There is no hard and fast rule about which variable is the response and which is the predictor.

Bivariate Data

Often, one variable is easier to measure than the other.

Bivariate Data

Often, one variable is easier to measure than the other.

For example, suppose we are interested in determining the percentage of the households in a zip code whose annual income puts them below the poverty line.

Bivariate Data

Often, one variable is easier to measure than the other.

For example, suppose we are interested in determining the percentage of the households in a zip code whose annual income puts them below the poverty line.

In this case, we would consider this percentage to be the response variable.

Bivariate Data

Often, one variable is easier to measure than the other.

For example, suppose we are interested in determining the percentage of the households in a zip code whose annual income puts them below the poverty line.

In this case, we would consider this percentage to be the response variable.

As a predictor variable, we might consider the number of pawn shops and bail bondsmen in the zip code.

Bivariate Data

Often, one variable is easier to measure than the other.

For example, suppose we are interested in determining the percentage of the households in a zip code whose annual income puts them below the poverty line.

In this case, we would consider this percentage to be the response variable.

As a predictor variable, we might consider the number of pawn shops and bail bondsmen in the zip code.

Note that we cannot assume a causal relationship between the response and predictor.

Bivariate Data

Often, one variable is easier to measure than the other.

For example, suppose we are interested in determining the percentage of the households in a zip code whose annual income puts them below the poverty line.

In this case, we would consider this percentage to be the response variable.

As a predictor variable, we might consider the number of pawn shops and bail bondsmen in the zip code.

Note that we cannot assume a causal relationship between the response and predictor.

While pawn shops and bail bondsmen are associated with poverty, they do not cause it.

Bivariate Data

Bivariate Data

For example, suppose we are interested in determining the percentage of the households in a zip code whose annual income puts them below the poverty line.

Bivariate Data

For example, suppose we are interested in determining the percentage of the households in a zip code whose annual income puts them below the poverty line.

In this case, we would consider this percentage to be the response variable.

Bivariate Data

For example, suppose we are interested in determining the percentage of the households in a zip code whose annual income puts them below the poverty line.

In this case, we would consider this percentage to be the response variable.

As a predictor variable, we might consider the number of pawn shops and bail bondsmen in the zip code.

Bivariate Data

For example, suppose we are interested in determining the percentage of the households in a zip code whose annual income puts them below the poverty line.

In this case, we would consider this percentage to be the response variable.

As a predictor variable, we might consider the number of pawn shops and bail bondsmen in the zip code.

Note that we cannot assume a causal relationship between the response and predictor.

Bivariate Data

For example, suppose we are interested in determining the percentage of the households in a zip code whose annual income puts them below the poverty line.

In this case, we would consider this percentage to be the response variable.

As a predictor variable, we might consider the number of pawn shops and bail bondsmen in the zip code.

Note that we cannot assume a causal relationship between the response and predictor.

While pawn shops are associated with poverty, they do not cause it.

Bivariate Data

Bivariate Data

Bivariate Data

In this case, we would consider this percentage to be the response variable.

Bivariate Data

In this case, we would consider this percentage to be the response variable.

As a predictor variable, we might consider the number of pawn shops and bail bondsmen in the zip code.

Bivariate Data

In this case, we would consider this percentage to be the response variable.

As a predictor variable, we might consider the number of pawn shops and bail bondsmen in the zip code.

Note that we cannot assume a causal relationship between the response and predictor.

Bivariate Data

In this case, we would consider this percentage to be the response variable.

As a predictor variable, we might consider the number of pawn shops and bail bondsmen in the zip code.

Note that we cannot assume a causal relationship between the response and predictor.

While pawn shops are associated with poverty, they do not cause it.

Correlation Coefficient

The **correlation coefficient** or Pearson correlation coefficient r is a statistic that measures the **linear** association between two variables.

Correlation Coefficient

The **correlation coefficient** or Pearson correlation coefficient r is a statistic that measures the **linear** association between two variables.

If one variable is exactly proportional to the other, the correlation coefficient will be either -1 or 1.

If there is no linear association at all between the variables, r will be zero (or close to zero)

Correlation Coefficient

The **correlation coefficient** or Pearson correlation coefficient r is a statistic that measures the **linear** association between two variables.

If one variable is exactly proportional to the other, the correlation coefficient will be either -1 or 1.

If there is no linear association at all between the variables, r will be zero (or close to zero)

In all cases r is between -1 and 1, inclusive.

Correlation Coefficient

The **correlation coefficient** or Pearson correlation coefficient r is a statistic that measures the **linear** association between two variables.

If one variable is exactly proportional to the other, the correlation coefficient will be either -1 or 1.

If there is no linear association at all between the variables, r will be zero (or close to zero)

In all cases r is between -1 and 1, inclusive.

Note that r measures only *linear* association.

Correlation Coefficient

The **correlation coefficient** or Pearson correlation coefficient r is a statistic that measures the **linear** association between two variables.

If one variable is exactly proportional to the other, the correlation coefficient will be either -1 or 1.

If there is no linear association at all between the variables, r will be zero (or close to zero)

In all cases r is between -1 and 1, inclusive.

Note that r measures only *linear* association.

It's possible two variables to have a zero correlation and yet be completely dependent on one another - just not linearly.